Scilight

Translational Insights
https://www.sciltp.com/journals/ti

Supplementary Materials

Efficient Adaptation of Chemical Language Models for Molecular Property Prediction

Kavya Agrawal [1,†], Harsha Harod [1,†], Ines Simeone [2], Michele Ceccarelli [3], Halima Bensmail [4], Sukrit Gupta [1] and Raghvendra Mall [4,5,*]

[1] Center for Applied Research in Data Science, Indian Institute of Technology Ropar, Punjab 140001, India
[2] Humanitas Research, 20089 Rozzano, Italy
[3] Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Coral Gables, FL 33136, USA
[4] Qatar Center for Artificial Intelligence, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha P.O. Box 5825, Qatar
[5] Biotechnology Research Center, Technology Innovation Institute, Abu Dhabi P.O. Box 9639, United Arab Emirates
* Correspondence: ramall@hbku.edu.qa or raghvendramall@ieee.org
† These authors contributed equally to this work.

How To Cite: Agrawal, K.; Harod, H.; Simeone, I.; et al. Efficient Adaptation of Chemical Language Models for Molecular Property Prediction. *Translational Insights* **2026**, *1*(1), 13. https://doi.org/10.53941/ti.2026.100013

1. Zero-Shot vs Full finetuning

The vanilla full finetuning of a CLM updates all the parameters of the pretrained model using task-specific data. It is similar to training a chemist in detailed lab techniques for a specific experiment [1,2]. This method works well in tasks that require high precision and where there is an abundance of labeled data [1]. It is particularly effective in capturing domain-specific patterns, such as improving the similarity of drugs in generated molecules [2]. However, it requires a large number of labeled examples and can easily be overfitted to small or narrow datasets [3]. Furthermore, there is a risk of catastrophic forgetting, where the model loses the knowledge acquired during pretraining [4]. These challenges are amplified as the size of the pretrained language model increases.

On the other hand, zero-shot learning uses prompts on / embeddings from the pretrained model to perform new tasks without changing their core parameters. It is analogous to giving a chemist a checklist for a new experiment rather than retraining them from scratch [5]. This method performs well when predicting results for new assays [5], and helps keep general molecular knowledge of pretraining intact [3]. However, it can struggle with tasks that are very different from the ones the generalist model learnt during pretraining [5] as there is no-backpropagation of losses.

Zero-shot models are cheap and have low latency, but are less accurate for specific tasks [5]. Full finetuning is computationally expensive and resource intensive as it creates models that can generalize for a specific task but cannot be easily re-used [1,2].

2. Low Rank Adaptation (LoRA)

The LoRA technique works through matrix decomposition, where the size of the matrix is reduced by splitting it into several other matrices to accelerate computation, the rationale being to capture the intrinsic dimensionality of language models for an effective finetuning [6].

The LoRA method introduces two new trainable low-rank matrices, often called adapters, consisting of a downward projection matrix ($\boldsymbol{A}$) and an upward projection matrix ($\boldsymbol{B}$). During LoRA finetuning, only these low-rank matrices are trained, and all the weights of the CLM are frozen.

Compared to the full finetuning method, it is computationally efficient to train these additional parameters. After training, we package up the LoRA adapter as a separate model file that can then be plugged-in to the base CLM. A fully finetuned model can be tens of gigabytes (Gb) in size, while these adapters are usually just a few megabytes (Mb). This makes it a lot easier to distribute, and providing finetuned inference with LoRA only adds milliseconds to the total inference time.

Formally, the original pre-trained weights matrices of the CLM are denoted as $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, while the two adapter matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$, are of sizes $\boldsymbol{A} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r\times d}$, where $r \ll d$. The rank parameter r is a hyper-parameter representing the intrinsic low dimensionality of the weight matrix. A higher rank means that more parameters need to be trained, allowing us to balance computation time and performance [7]. The matrix $\boldsymbol{A}$ is

initialized with normally distributed random values of zero mean and unit variance while $\boldsymbol{B}$ is initialized with zeroes. This ensures that they do not significantly alter the original $\boldsymbol{W}$ at the start of the training. The matrices are then fed with the same input as $\boldsymbol{W}$ and their output is added with the output of the CLM. The update of weight matrix is represented as:

$$\Delta \boldsymbol{W} = \boldsymbol{A} \cdot \boldsymbol{B} \tag{S1}$$

Here the LoRA technique targets the attention block in the transformer architecture of the CLM and processes low-rank matrices in parallel with the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ matrices of the attention layer.

3. EffiChem Trainer

An essential component of the EffiChem architecture is the custom trainer designed to address class imbalance issues prevalent in molecular classification tasks and to add compute metrics such F1 score, Accuracy, MCC, Precision and Recall.

Loss Function: To address the class imbalance, we modified the 'Trainer' class from the `Huggingface` library and introduced two types of loss functions:

(1) Weighted Loss

To address class imbalance in both binary and multi-class classification tasks, we employed a weighted cross-entropy loss. The weight for each class was computed based on the inverse frequency of that class in the training data. For weight calculation, inverse class frequency (i.e., $w_i = 1 - f_i$, where f_i was i^{th} class frequency) was assigned to give higher weights to the minority class. For example if 90% samples belong class to 0 and 10% fall in class 1, then class weights assigned were $[0.1, 0.9]$.

The final weighted cross-entropy loss for binary classification instance was: $\mathscr{L}_{\text{binary}} = -\sum_{i=0}^{1} w_i \cdot y_i \cdot \log(\hat{y}_i)$, where $y_i \in \{0, 1\}$ was the ground truth label, $\hat{y}_i \in [0, 1]$ was the predicted probability for class i and w_i was the weight for class i. Similar approach was extended to the multi-class dataset. For example, classes with frequencies [0.6, 0.3, 0.1], were assigned weights = [0.4, 0.7, 0.9]. The intuition was to prioritize rare classes. Thus, the weighted cross-entropy loss became:

$$\mathscr{L}_{\text{multi}} = -\sum_{i=0}^{C-1} w_i \cdot y_i \cdot \log(\hat{y}_i) \tag{S2}$$

where $y_i \in \{0, 1\}$ was the one-hot encoded ground truth label, $\hat{y}_i \in [0, 1]$ was the predicted probability for class i and w_i was the class weight.

These formulations allow the model to assign more importance to under-represented classes, which helps mitigate performance bias due to class imbalance. The computed weights are incorporated into the weighted cross-entropy loss function, improving generalization across all classes.

(2) Focal Loss

To further address class imbalance and focus training on hard-to-classify examples, we used the *focal loss*. Unlike standard cross-entropy loss, focal loss reduces the relative loss for well-classified examples and emphasizes learning from hard negatives. This was particularly useful for imbalanced classes. Given C be the number of classes, and $\hat{p}_i$ be the predicted probability for the true class i. The focal loss for a single instance was defined as:

$$\mathscr{L}_{\text{focal}} = -\alpha \cdot (1 - \hat{p}_i)^{\gamma} \cdot \log(\hat{p}_i)$$

where $\alpha \in [0, 1]$ was a balancing factor, $\gamma \geq 0$ was the focusing parameter and $\hat{p}_i = \text{softmax}(\mathbf{z})_i$, with $\mathbf{z}$ being the output logits.

Although originally proposed for multi-class problems, we used a generalized multi-class focal loss implementation that works with any number of classes C. When using the `Huggingface` library's `AutoModelForSequenceClassification` with `num_labels` parameter set to 2, the model outputs two logits, effectively treating binary classification as a special case of multi-class classification. Thus, our focal loss function naturally extends to both binary and multi-class classification tasks without requiring separate implementations.

4. Experimental Results

In this section, we undertake a comprehensive ablation study. We compare the performance of tree-based models built on physico-chemical (PC) features (see Table S1), base embeddings (i.e. zero-shot embedding

representation from CLMs, see Table S2), PC + base embeddings, embeddings generated from finetuned CLMs, i.e. finetuned embeddings, and PC + finetuned embeddings. We compare these models with end-to-end LoRA finetuned CLM models (EffiChem architecture) optimized for classification task.

4.1. BBBP Task Performance

We first perform a comprehensive set of comparisons for the BBBP dataset.

4.1.1. Top Performing Methods

Our comprehensive evaluation of the BBBP dataset reveals that end-to-end (E2E) LoRA finetuning of CLMs yields the strongest overall performance. The top three methods based on average performance in all evaluation metrics were:

- E2E LoRA MolFormer-XL-10pct-both (weight Loss) - The finetuning of MolFormer with weighted loss by E2E LoRA achieved the best overall performance, with an MCC of 0.797 and a macro F1 of 0.898. This method attained the highest Accuracy of 0.906 and F1 micro of 0.906 among all the evaluated approaches, demonstrating that direct optimization of the model for the classification task provides superior discriminative capability (see Table S3).
- E2E LoRA MolFormer-XL-10pct-both (Focal Loss) - The focal loss variant of E2E LoRA finetuning ranked second, achieving an MCC of 0.778 and F1 macro of 0.863. With an AUC of 0.936 and Accuracy of 0.896, this approach confirms the effectiveness of the LoRA architecture for BBBP property prediction.
- FT Embed MolFormer-XL-10pct-both (Focal Loss) with XGBoost - Finetuned MolFormer embeddings combined with XGBoost using focal loss ranked third, achieving an MCC of 0.787 and F1 macro of 0.888. The Accuracy of 0.901 further validates the benefit of task-specific embedding adaptation when used with gradient boosting classifiers.

4.1.2. Performance Trends Across Feature Modalities

A clear trend of performance improvement emerges when progressing from simpler to more sophisticated feature representations as depicted in Tables S2 and S3.

- Physico-chemical (PC) features alone: The baseline using only PC features with tree-based models achieved moderate performance (best MCC: 0.732, F1 macro: 0.858 with LightGBM/CatBoost).
- Base CLM embeddings: Zero-shot embeddings from pretrained CLMs showed mixed results. ChemBERTa-77M MLM with XGBoost achieved MCC of 0.683 and F1 macro of 0.818, while MolFormer-XL with CatBoost reached MCC of 0.765 and F1 macro of 0.876, demonstrating that larger, more capable foundation models provide superior molecular representations.
- PC + Base embeddings: Combining PC features with base embeddings did not yield improvements over either modality alone. PC + ChemBERTa-77M MTR with CatBoost achieved MCC of 0.754 and F1 macro of 0.869, which could not surpass the optimal performance of base embeddings alone.
- Finetuned (FT) embeddings: Task-specific finetuning substantially improved performance. FT MolFormer embeddings with XGBoost (Focal Loss) achieved MCC of 0.787 and F1 macro of 0.888, while FT ChemBERTa-77M MLM with LightGBM (Focal Loss) reached MCC of 0.776 and F1 macro of 0.882.
- PC + FT embeddings: The combination of PC features with finetuned embeddings produced strong results but did not surpass end-to-end approaches. PC + FT MolFormer (Weighted Loss) with CatBoost achieved MCC of 0.773 and F1 macro of 0.884.
- End-to-end LoRA finetuning: Direct optimization through LoRA achieved the best performance overall. E2E LoRA MolFormer (Weighted Loss) attained the highest MCC of 0.797, F1 macro of 0.898, and Accuracy of 0.906 observed across all methods, establishing it as the optimal approach for the BBBP task.

These results demonstrate that (1) MolFormer consistently outperforms ChemBERTa across all experimental conditions; (2) E2E LoRA finetuning provides the best performance by directly optimizing the model for the downstream task (see Table S3); and (3) while combining engineered physico-chemical features with learned representations offers benefits, the end-to-end approach ultimately delivers superior performance using LoRA adapter by reducing the over-parameterization burden of MolFormer, highlighting the efficacy of EffiChem architecture.

4.2. Flavor Task Performance

We perform a comprehensive set of comparisons for the Flavor (FART) dataset.

Table S1. RDKit-derived molecular descriptors, 'networkx' based graphical features, Morgan and MACCS fingerprints obtained from RDKit, together represent the set of engineered features.

Feature Name	Feature Count	Feature Type
MaxAbsEStateIndex	1	Descriptor
MinAbsEStateIndex	1	Descriptor
MinEStateIndex	1	Descriptor
qed	1	Descriptor
SPS	1	Descriptor
MolWt	1	Physico-chemical
FpDensityMorgan1	1	Descriptor
AvgIpc	1	Descriptor
BalabanJ	1	Descriptor
Ipc	1	Descriptor
Kappa2	1	Descriptor
PEOE_VSA	14	Descriptor
SMR_VSA	10	Descriptor
SlogP_VSA	12	Descriptor
TPSA	1	Descriptor
EState_VSA	11	Descriptor
VSA_Estate	10	Descriptor
NHOHCount	1	Physico-chemical
NumAliphaticCarbocycles	1	Physico-chemical
NumAliphaticHeterocycles	1	Physico-chemical
NumAliphaticRings	1	Physico-chemical
NumAmideBonds	1	Physico-chemical
NumAromaticHeterocycles	1	Physico-chemical
NumAtomStereoCenters	1	Physico-chemical
NumBridgeheadAtoms	1	Physico-chemical
NumHAcceptors	1	Physico-chemical
NumHeteroatoms	1	Physico-chemical
NumHeterocycles	1	Physico-chemical
NumRotatableBonds	1	Physico-chemical
NumSaturatedCarbocycles	1	Physico-chemical
NumSaturatedHeterocycles	1	Physico-chemical
NumSaturatedRings	1	Physico-chemical
NumSpiroAtoms	1	Physico-chemical
NumUnspecifiedAtomStereoCenters	1	Physico-chemical
RingCount	1	Physico-chemical
MolLogP	1	Physico-chemical
Fraction of Functional Groups	62	Physico-chemical
graph_diameter	1	Graphical
avg_shortest_path	1	Graphical
num_cycles	1	Graphical
num_chains	1	Graphical
clustering_coefficients	1	Graphical
avg_degree_centrality	1	Graphical
avg_eigen_centrality	1	Graphical
avg_betweenness_centrality	1	Graphical
avg_load_centrality	1	Graphical
wiener_index	1	Graphical
max_degree	1	Graphical
closeness_mean	1	Graphical
katz_centrality_std	1	Graphical
betweenness_mean	1	Graphical
betweenness_std	1	Graphical
eigenvector_mean	1	Graphical
rings	5	Graphical
heteroatom_ratio	1	Graphical
num_aromatic_rings	1	Graphical
num_non_aromatic_rings	1	Graphical
average_atoms (C, O, S, N)	4	Graphical
num_single_bonds	1	Graphical
num_double_bonds	1	Graphical
Morgan	128	Fingerprint
MACCS	167	Fingerprint
Total	473	Combined

Table S2. Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results highlighted in bold.

Modality	CLM	ML-Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
PC	-	XGBoost	0.938	0.875	0.858	0.875	0.728	0.886	0.844
PC	-	LightGBM	0.931	0.875	0.856	0.875	0.732	0.896	0.838
PC	-	CatBoost	0.931	0.875	0.856	0.875	0.732	0.896	0.838
Base	ChemBERTa-77M MLM	XGBoost	0.889	0.849	0.818	0.849	0.683	0.894	0.796
		LightGBM	0.882	0.833	0.798	0.833	0.646	0.876	0.778
		CatBoost	0.884	0.823	0.790	0.823	0.614	0.846	0.772
Base	ChemBERTa-77M MTR	XGBoost	0.917	0.849	0.828	0.849	0.669	0.856	0.814
		LightGBM	0.879	0.792	0.762	0.792	0.537	0.787	0.751
		CatBoost	0.914	0.849	0.829	0.849	0.668	0.852	0.817
Base	MolFormer-XL-10pct-both	XGBoost	0.893	0.849	0.831	0.849	0.668	0.849	0.820
		LightGBM	0.920	0.844	0.817	0.844	0.662	0.868	0.798
		CatBoost	0.915	**0.891**	**0.876**	**0.891**	**0.765**	**0.907**	**0.859**
PC + Base	ChemBERTa-77M MLM	XGBoost	0.922	0.880	0.864	0.880	0.741	0.894	0.848
		LightGBM	0.933	0.880	0.865	0.880	0.739	0.890	0.851
		CatBoost	0.913	0.875	0.857	0.875	0.730	0.891	0.841
PC + Base	ChemBERTa-77M MTR	XGBoost	0.928	0.870	0.849	0.870	0.721	0.893	0.831
		LightGBM	0.934	0.880	0.862	0.880	0.743	0.900	0.845
		CatBoost	0.929	0.885	0.869	0.885	0.754	0.903	0.852
PC + Base	MolFormer-XL-10pct-both	XGBoost	**0.940**	0.870	0.851	0.870	0.718	0.887	0.834
		LightGBM	0.927	0.870	0.851	0.870	0.718	0.887	0.834
		CatBoost	0.920	0.859	0.840	0.859	0.693	0.869	0.825

Table S3. Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold.

Modality	CLM	Loss Function	ML Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
E2E LoRA	ChemBERTa-77M MLM	Focal Loss	-	0.923	0.88	0.851	0.865	0.739	0.889	0.851
		Weighted Loss	-	0.929	0.875	0.862	0.875	0.727	0.871	0.856
E2E LoRA	ChemBERTa-77M MTR	Focal Loss	-	0.921	0.87	0.843	0.854	0.715	0.873	0.843
		Weighted Loss	-	0.906	0.854	0.824	0.854	0.68	0.856	0.824
E2E LoRA	MolFormer-XL-10pct-both	Focal Loss	-	0.936	0.896	0.863	0.881	0.778	**0.916**	0.863
		Weighted Loss	-	0.948	**0.906**	**0.898**	**0.906**	**0.797**	0.901	**0.896**
FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.909	0.875	0.857	0.875	0.730	0.89	0.841
			LightGBM	0.905	0.896	0.882	0.896	0.776	0.911	0.866
			CatBoost	0.925	0.885	0.871	0.885	0.751	0.893	0.858
FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.921	0.875	0.855	0.875	0.734	0.903	0.835
			LightGBM	0.915	0.870	0.853	0.87	0.716	0.877	0.840
			CatBoost	0.927	0.870	0.853	0.87	0.716	0.877	0.840
FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.923	0.854	0.833	0.854	0.681	0.865	0.818
			LightGBM	0.902	0.865	0.846	0.865	0.704	0.873	0.833
			CatBoost	0.926	0.854	0.838	0.854	0.68	0.85	0.830
FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.905	0.849	0.829	0.849	0.668	0.852	0.817
			LightGBM	0.887	0.859	0.842	0.859	0.692	0.861	0.831
			CatBoost	0.911	0.865	0.849	0.865	0.703	0.865	0.839
FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.935	0.901	0.888	0.901	0.787	0.914	0.873
			LightGBM	0.935	0.885	0.868	0.885	0.756	0.909	0.849
			CatBoost	0.936	0.885	0.872	0.885	0.75	0.889	0.861
FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.948	0.870	0.850	0.870	0.718	0.887	0.834
			LightGBM	0.945	0.875	0.857	0.875	0.730	0.890	0.841
			CatBoost	0.949	0.88	0.864	0.880	0.741	0.894	0.848
PC + FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.917	0.875	0.857	0.875	0.730	0.89	0.841
			LightGBM	0.888	0.891	0.875	0.891	0.765	0.907	0.859
			CatBoost	0.928	0.875	0.860	0.875	0.727	0.877	0.850
PC + FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.923	0.859	0.840	0.859	0.693	0.869	0.825
			LightGBM	0.920	0.870	0.853	0.870	0.716	0.877	0.840
			CatBoost	0.928	0.870	0.850	0.870	0.718	0.887	0.834
PC + FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.923	0.859	0.841	0.859	0.692	0.865	0.828
			LightGBM	0.909	0.844	0.823	0.844	0.657	0.848	0.810
			CatBoost	0.928	0.854	0.838	0.854	0.680	0.850	0.830
PC + FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.897	0.854	0.833	0.854	0.681	0.865	0.818
			LightGBM	0.886	0.849	0.827	0.849	0.670	0.861	0.811
			CatBoost	0.914	0.859	0.844	0.859	0.692	0.858	0.835
PC + FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.937	0.885	0.871	0.885	0.751	0.893	0.858
			LightGBM	0.934	0.901	0.887	0.901	0.789	0.920	0.870
			CatBoost	0.931	0.901	0.888	0.901	0.787	0.914	0.873
PC + FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.949	0.875	0.856	0.875	0.732	0.896	0.838
			LightGBM	0.945	0.885	0.869	0.885	0.754	0.903	0.852
			CatBoost	**0.951**	0.896	0.884	0.896	0.773	0.902	0.872

4.2.1. Top Performing Methods

Our comprehensive evaluation on the Flavor dataset reveals that E2E LoRA finetuning of CLMs yields the strongest overall performance (see Table S4). The top three methods based on average performance across all

evaluation metrics are:

- E2E LoRA MolFormer-XL-10pct-both (Weighted Loss) - End-to-end LoRA finetuning of MolFormer with weighted loss achieved the best overall performance, with an MCC of 0.801 and F1 macro of 0.809. This method attained an AUC of 0.975 and Accuracy of 0.892, with balanced Precision of 0.802 and Recall of 0.817, demonstrating robust classification across all flavor categories.
- E2E LoRA ChemBERTa-77M MTR (Weighted Loss) - The weighted loss variant of end-to-end LoRA finetuning with ChemBERTa-77M MTR ranked second, achieving an MCC of 0.802 and F1 macro of 0.780. With an AUC of 0.964 and Accuracy of 0.890, this approach showed well-balanced Precision of 0.780 and Recall of 0.786, confirming the effectiveness of the LoRA architecture even with smaller language models.
- E2E LoRA MolFormer-XL-10pct-both (Focal Loss) - The focal loss variant of end-to-end LoRA finetuning with MolFormer ranked third, achieving an MCC of 0.792 and F1 macro of 0.789. With an AUC of 0.971 and Accuracy of 0.889, this method achieved Precision of 0.820 and Recall of 0.770, indicating a slight bias toward precision over recall.

4.2.2. Performance Trends Across Feature Modalities

A clear trend of performance improvement emerges when progressing from simpler to more sophisticated feature representations as depicted in Tables S5 and S4:

- Physico-chemical (PC) features alone: The baseline using only PC features with tree-based models achieved strong performance. CatBoost achieved the best balance with MCC of 0.786, F1 macro of 0.776, Precision of 0.732, and Recall of 0.849, notably the highest recall among PC-only methods, though with lower precision.
- Base CLM embeddings: Zero-shot embeddings from pretrained CLMs showed variable results. ChemBERTa-77M MLM embeddings performed poorly (MCC: 0.698–0.746, F1 macro: 0.624–0.634), with low Precision (0.591–0.626) and Recall (0.643–0.677). In contrast, ChemBERTa-77M MTR with LightGBM achieved MCC of 0.795, F1 macro of 0.767, Precision of 0.749, and Recall of 0.790, approaching PC feature performance. MolFormer-XL with XGBoost achieved MCC of 0.761 and F1 macro of 0.743, but with notably higher Precision (0.822) and lower Recall (0.737) as illustrated in Table S5.
- PC + Base embeddings: Combining PC features with base embeddings yielded improvements. PC + ChemBERTa-77M MTR with CatBoost achieved MCC of 0.798, F1 macro of 0.801, Precision of 0.782, and Recall of 0.824 demonstrating that the combination captures complementary information. However, PC + MolFormer combinations did not consistently improve over PC alone (see Table S5).
- Finetuned (FT) embeddings: Task-specific finetuning showed strong performance across most configurations. FT MolFormer embeddings with LightGBM (Weighted Loss) achieved MCC of 0.787, F1 macro of 0.783, Precision of 0.800, and Recall of 0.778. FT ChemBERTa-77M MLM with CatBoost (Weighted Loss) reached MCC of 0.783, F1 macro of 0.726, but with lower Precision (0.695) and higher Recall (0.779).
- PC + FT embeddings: The combination of PC features with finetuned embeddings produced competitive but not superior results. PC + FT MolFormer (Weighted Loss) with LightGBM achieved MCC of 0.789, F1 macro of 0.798, with the highest Precision among combined approaches (0.852) but lower Recall (0.779). PC + FT ChemBERTa-77M MTR (Weighted Loss) with LightGBM achieved MCC of 0.800, F1 macro of 0.742, Precision of 0.708, and Recall of 0.794 as highlighted in Table S4.
- End-to-end LoRA finetuning: Direct optimization through LoRA achieved the best performance overall. E2E LoRA MolFormer (Weighted Loss) attained the highest F1 macro of 0.809, MCC of 0.801, with well-balanced Precision (0.802) and Recall (0.817). E2E LoRA ChemBERTa-77M MTR (Weighted Loss) achieved competitive MCC of 0.802 and F1 macro of 0.780, with balanced Precision (0.780) and Recall (0.786).

4.2.3. Key Observations on Precision-Recall Trade-Offs

The Flavor task exhibits interesting precision-recall dynamics across methods:

- PC features with CatBoost favor recall (0.849) over precision (0.732), potentially useful when missing flavor compounds is costly.
- MolFormer-based methods tend toward higher precision, with base MolFormer + XGBoost achieving Precision of 0.822 but Recall of only 0.737.
- End-to-end LoRA methods achieve the best balance, with E2E LoRA MolFormer (Weighted Loss) showing nearly equal Precision (0.802) and Recall (0.817).
- Focal loss variants generally emphasize precision over recall, while weighted loss variants achieve better balance between the two metrics as observed from Table S4.

Table S4. Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold.

Modality	CLM	Loss Function	ML Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
E2E LoRA	ChemBERTa-77M MLM	Focal Loss	-	0.959	0.881	0.691	0.881	0.772	0.714	0.674
		Weighted Loss	-	**0.975**	0.882	0.734	0.882	0.785	0.732	0.738
E2E LoRA	ChemBERTa-77M MTR	Focal Loss	-	0.963	0.891	0.728	0.891	0.795	0.872	0.702
		Weighted Loss	-	0.964	0.890	0.780	0.890	**0.802**	0.780	0.786
E2E LoRA	MolFormer-XL-10pct-both	Focal Loss	-	0.971	0.889	0.789	0.889	0.792	0.820	0.770
		Weighted Loss	-	**0.975**	**0.892**	**0.809**	**0.892**	0.801	0.802	**0.817**
FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.962	0.872	0.731	0.872	0.773	0.733	0.741
			LightGBM	0.963	0.873	0.727	0.873	0.775	0.717	0.742
			CatBoost	0.963	0.872	0.724	0.872	0.772	0.715	0.738
FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.971	0.877	0.717	0.877	0.783	0.693	0.750
			LightGBM	0.971	0.877	0.739	0.877	0.783	0.709	0.781
			CatBoost	0.973	0.877	0.726	0.877	0.783	0.695	0.779
FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.953	0.881	0.738	0.881	0.793	0.724	0.759
			LightGBM	0.956	0.877	0.716	0.877	0.786	0.687	0.756
			CatBoost	0.969	0.878	0.719	0.878	0.786	0.693	0.754
FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.962	0.878	0.714	0.878	0.787	0.687	0.753
			LightGBM	0.964	0.879	0.742	0.879	0.789	0.711	0.786
			CatBoost	0.972	0.879	0.731	0.879	0.789	0.697	0.788
FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.960	0.875	0.764	0.875	0.776	0.763	0.771
			LightGBM	0.950	0.879	0.766	0.879	0.779	0.770	0.766
			CatBoost	0.966	0.874	0.764	0.874	0.778	0.758	0.777
FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.969	0.879	0.749	0.879	0.785	0.777	0.748
			LightGBM	0.970	0.882	0.783	0.882	0.787	0.800	0.778
			CatBoost	0.957	0.879	0.741	0.879	0.787	0.710	0.783
PC + FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.925	0.775	0.624	0.775	0.655	0.608	0.678
			LightGBM	0.939	0.844	0.700	0.844	0.732	0.689	0.724
			CatBoost	0.961	0.872	0.723	0.872	0.771	0.715	0.736
PC + FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.954	0.837	0.682	0.837	0.730	0.650	0.739
			LightGBM	0.947	0.756	0.644	0.756	0.641	0.615	0.748
			CatBoost	0.971	0.879	0.736	0.879	0.786	0.707	0.780
PC + FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.919	0.756	0.633	0.756	0.643	0.606	0.717
			LightGBM	0.935	0.843	0.709	0.843	0.739	0.693	0.744
			CatBoost	0.965	0.878	0.727	0.878	0.785	0.712	0.747
PC + FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.933	0.766	0.648	0.766	0.656	0.625	0.723
			LightGBM	0.948	0.886	0.742	0.886	0.800	0.708	0.794
			CatBoost	0.964	0.885	0.754	0.885	0.798	0.728	0.788
PC + FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.960	0.866	0.759	0.866	0.763	0.754	0.770
			LightGBM	0.961	0.881	0.781	0.881	0.783	0.800	0.774
			CatBoost	0.964	0.874	0.775	0.874	0.778	0.788	0.778
PC + FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.965	0.882	0.785	0.882	0.789	0.798	0.786
			LightGBM	0.967	0.883	0.798	0.883	0.789	**0.852**	0.779
			CatBoost	0.958	0.888	0.772	0.888	0.801	0.756	0.793

Table S5. Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results are highlighted in bold.

Modality	CLM	ML-Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
PC	-	XGBoost	**0.974**	**0.888**	0.751	**0.888**	**0.803**	0.721	0.793
PC	-	LightGBM	0.970	0.886	0.768	0.886	0.800	0.752	0.790
PC	-	CatBoost	0.965	0.878	0.776	0.878	0.786	0.732	**0.849**
Base	ChemBERTa-77M MLM	XGBoost	0.916	0.862	0.634	0.862	0.745	0.626	0.643
		LightGBM	0.912	0.857	0.632	0.857	0.746	0.614	0.655
		CatBoost	0.874	0.819	0.624	0.819	0.698	0.591	0.677
Base	ChemBERTa-77M MTR	XGBoost	0.972	0.874	0.756	0.874	0.779	0.737	0.781
		LightGBM	0.973	0.884	0.767	0.884	0.795	0.749	0.790
		CatBoost	0.967	0.877	0.759	0.877	0.785	0.740	0.784
Base	MolFormer-XL-10pct-both	XGBoost	0.960	0.863	0.743	0.863	0.761	0.822	0.737
		LightGBM	0.965	0.870	0.706	0.870	0.770	**0.830**	0.707
		CatBoost	0.956	0.860	0.699	0.860	0.758	0.671	0.737
PC + Base	ChemBERTa-77M MLM	XGBoost	0.958	0.882	0.751	0.882	0.793	0.776	0.753
		LightGBM	0.957	0.883	0.712	0.883	0.793	0.744	0.719
		CatBoost	0.947	0.872	0.732	0.872	0.776	0.700	0.776
PC + Base	ChemBERTa-77M MTR	XGBoost	0.972	0.887	0.781	0.887	0.802	0.772	0.796
		LightGBM	0.968	0.886	0.779	0.886	0.800	0.770	0.795
		CatBoost	0.965	0.885	**0.801**	0.885	0.798	0.782	0.824
PC + Base	MolFormer-XL-10pct-both	XGBoost	0.968	0.881	0.706	0.881	0.789	0.709	0.719
		LightGBM	0.966	0.879	0.741	0.879	0.786	0.714	0.776
		CatBoost	0.957	0.874	0.721	0.874	0.779	0.702	0.747

These results demonstrate that (1) E2E LoRA finetuning provides the best overall performance with balanced precision and recall (see Table S4); (2) MolFormer and ChemBERTa-77M MTR both serve as effective base models for this multi-class task (see Table S5); and (3) the choice of loss function (focal vs. weighted) influences the

precision-recall trade-off, with weighted loss generally achieving better balance for flavor prediction (see Table S4).

4.3. ClinTox Task Performance

We perform a comprehensive comparisons on the ClinTox dataset for toxicity prediction task.

4.3.1. Top Performing Methods

Our comprehensive evaluation on the ClinTox dataset for toxicity prediction reveals that combining physico-chemical features with base CLM embeddings yields the strongest overall performance as observed from Tables S6 and S7. Notably, this task exhibits exceptionally high performance across most methods, with several achieving near-perfect AUC scores. The top three methods based on average performance across all evaluation metrics are:

- PC + Base MolFormer-XL-10pct-both with LightGBM - Combining PC features with base MolFormer embeddings using LightGBM ranked best, achieving an AUC of 0.999, MCC of 0.875, and F1 macro of 0.934. This method achieved the highest MCC among all approaches, with Accuracy of 0.986, Precision of 0.889, and Recall of 0.993, the highest recall observed across all methods.
- PC + Base MolFormer-XL-10pct-both with CatBoost - This combination ranked second with an AUC of 0.998, MCC of 0.850, and F1 macro of 0.925. The method achieved Accuracy of 0.986, with balanced Precision of 0.925 and Recall of 0.925, demonstrating robust and consistent toxicity prediction.
- PC + Base ChemBERTa-77M MLM with CatBoost - This method achieved the best AUC of 1.000, MCC of 0.827, and F1 macro of 0.906. The method attained Accuracy of 0.979, with Precision of 0.850 and Recall of 0.989, demonstrating exceptional ability to identify toxic compounds while maintaining high precision.

4.3.2. Performance Trends Across Feature Modalities

The ClinTox task presents a unique pattern compared to other molecular property prediction tasks, with base embeddings and their combinations outperforming finetuned approaches as observed from Tables S6 and S7:

- Physico-chemical (PC) features alone: The baseline using only PC features with tree-based models showed variable performance. LightGBM achieved MCC of 0.700, F1 macro of 0.850, with balanced Precision (0.850) and Recall (0.850). However, XGBoost showed lower precision (0.673) despite high recall (0.888), and CatBoost exhibited the similar pattern with very high Recall (0.912) but lower Precision (0.615).
- Base CLM embeddings: Zero-shot embeddings from pretrained CLMs showed strong performance on this task. Base MolFormer-XL with LightGBM achieved the highest single-modality performance with MCC of 0.850, F1 macro of 0.925, Precision of 0.925, and Recall of 0.925. ChemBERTa-77M MLM with XGBoost achieved MCC of 0.761, F1 macro of 0.879, with high Precision (0.909) but lower Recall (0.854). ChemBERTa-77M MTR with LightGBM reached MCC of 0.791, F1 macro of 0.895, with Precision of 0.871 and Recall of 0.921.
- PC + Base embeddings: Combining PC features with base embeddings yielded the best overall results on this task. PC + ChemBERTa-77M MLM with CatBoost achieved a perfect AUC of 1.000, MCC of 0.827, F1 macro of 0.906, with Precision of 0.850 and exceptional Recall of 0.989. PC + MolFormer with LightGBM achieved the highest MCC of 0.875, F1 macro of 0.934, with Precision of 0.889 and the highest Recall of 0.993.
- Finetuned (FT) embeddings: Surprisingly, task-specific finetuning did not improve upon base embeddings for this task (see Table S7). FT MolFormer embeddings with XGBoost/LightGBM (Focal Loss) achieved MCC of 0.827, F1 macro of 0.906, Precision of 0.850, and Recall of 0.989. FT ChemBERTa-77M MLM with CatBoost (Focal Loss) reached MCC of 0.850, F1 macro of 0.925, with balanced Precision (0.925) and Recall (0.925), matching but not exceeding base embedding performance.
- PC + FT embeddings: The combination of PC features with finetuned embeddings produced results comparable and worse than PC + base embeddings. PC + FT ChemBERTa-77M MTR (Focal Loss) with CatBoost achieved MCC of 0.786, F1 macro of 0.881, Precision of 0.818, and Recall of 0.985. PC + FT MolFormer combinations generally achieved MCC around 0.717–0.750, F1 macro of $0.839 - 0.859$, underperforming when comapred to the PC + base embedding combinations.
- End-to-end LoRA finetuning: Direct optimization through LoRA did not achieve the best results on this task. E2E LoRA MolFormer (Focal Loss) attained MCC of 0.786, F1 macro of 0.881, Precision of 0.818, and Recall of 0.985. E2E LoRA ChemBERTa-77M MLM (Focal Loss) achieved MCC of 0.717, F1 macro of 0.839, with Precision of 0.769 and Recall of 0.978. While recall remains high, precision and MCC are lower than the best PC + base embedding methods. This can partly be attributed to the immense class imbalance in the ClinTox dataset making the learning difficult for E2E models and embeddings generated from the finetuned models.

Table S6. Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results are highlighted in bold

Modality	CLM	ML-Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
PC	-	XGBoost	0.975	0.916	0.727	0.916	0.518	0.673	0.888
PC	-	LightGBM	0.975	0.972	0.850	0.972	0.700	0.850	0.850
PC	-	CatBoost	0.986	0.832	0.636	0.832	0.431	0.615	0.912
Base	ChemBERTa-77M MLM	XGBoost	0.983	0.979	0.879	0.979	0.761	0.909	0.854
		LightGBM	0.990	0.972	0.850	0.972	0.700	0.850	0.850
		CatBoost	0.983	0.958	0.822	0.958	0.664	0.769	0.910
Base	ChemBERTa-77M MTR	XGBoost	0.988	0.965	0.844	0.965	0.700	0.796	0.914
		LightGBM	0.979	0.979	0.895	0.979	0.791	0.871	0.921
		CatBoost	0.987	0.965	0.844	0.965	0.700	0.796	0.914
Base	MolFormer-XL-10pct-both	XGBoost	0.994	0.979	0.895	0.979	0.791	0.871	0.921
		LightGBM	0.997	**0.986**	0.925	**0.986**	0.850	**0.925**	0.925
		CatBoost	0.998	0.979	0.895	0.979	0.791	0.871	0.921
PC + Base	ChemBERTa-77M MLM	XGBoost	0.998	0.972	0.881	0.972	0.786	0.818	0.985
		LightGBM	0.999	0.972	0.881	0.972	0.786	0.818	0.985
		CatBoost	**1.000**	0.979	0.906	0.979	0.827	0.850	0.989
PC + Base	ChemBERTa-77M MTR	XGBoost	0.984	0.972	0.881	0.972	0.786	0.818	0.985
		LightGBM	0.988	0.965	0.844	0.965	0.700	0.796	0.914
		CatBoost	0.990	0.965	0.844	0.965	0.700	0.796	0.914
PC + Base	MolFormer-XL-10pct-both	XGBoost	0.996	0.979	0.895	0.979	0.791	0.871	0.921
		LightGBM	0.999	**0.986**	**0.934**	**0.986**	**0.875**	0.889	**0.993**
		CatBoost	0.998	**0.986**	0.925	**0.986**	0.850	**0.925**	0.925

Table S7. Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold.

Modality	CLM	Loss Function	ML Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
E2E LoRA	ChemBERTa-77M MLM	Focal Loss	-	0.987	0.958	0.839	0.958	0.717	0.769	0.978
		Weighted Loss	-	0.988	0.951	0.820	0.951	0.689	0.750	0.974
E2E LoRA	ChemBERTa-77M MTR	Focal Loss	-	0.985	0.958	0.822	0.958	0.664	0.769	0.910
		Weighted Loss	-	0.987	0.965	0.859	0.965	0.750	0.792	0.982
E2E LoRA	MolFormer-XL-10pct-both	Focal Loss	-	0.993	0.972	0.881	0.972	0.786	0.818	0.985
		Weighted Loss	-	0.982	0.958	0.839	0.958	0.717	0.769	0.978
FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.988	0.958	0.822	0.958	0.664	0.769	0.910
			LightGBM	0.985	0.972	0.868	0.972	0.742	0.830	0.918
			CatBoost	0.987	0.986	0.925	0.986	0.850	0.925	0.925
FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.986	0.965	0.859	0.965	0.750	0.792	0.982
			LightGBM	0.993	0.972	0.881	0.972	0.786	0.818	0.985
			CatBoost	0.989	0.951	0.820	0.951	0.689	0.750	0.974
FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.986	0.951	0.820	0.951	0.689	0.750	0.974
			LightGBM	0.986	0.958	0.839	0.958	0.717	0.769	0.978
			CatBoost	**0.994**	0.958	0.839	0.958	0.717	0.769	0.978
FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.983	0.958	0.839	0.958	0.717	0.769	0.978
			LightGBM	0.981	0.958	0.839	0.958	0.717	0.769	0.978
			CatBoost	0.989	0.965	0.859	0.965	0.750	0.792	0.982
FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.987	**0.979**	**0.906**	**0.979**	**0.827**	**0.850**	**0.989**
			LightGBM	0.987	**0.979**	**0.906**	**0.979**	**0.827**	**0.850**	**0.989**
			CatBoost	0.987	0.965	0.859	0.965	0.750	0.792	0.982
FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.979	0.965	0.859	0.965	0.750	0.792	0.982
			LightGBM	0.981	0.965	0.859	0.965	0.750	0.792	0.982
			CatBoost	0.984	0.965	0.859	0.965	0.750	0.792	0.982
PC + FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.988	0.958	0.822	0.958	0.664	0.769	0.910
			LightGBM	0.988	0.951	0.803	0.951	0.633	0.746	0.907
			CatBoost	0.988	0.951	0.803	0.951	0.633	0.746	0.907
PC + FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.985	0.972	0.881	0.972	0.786	0.818	0.985
			LightGBM	0.989	0.972	0.881	0.972	0.786	0.818	0.985
			CatBoost	0.991	0.958	0.839	0.958	0.717	0.769	0.978
PC + FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.986	0.965	0.859	0.965	0.750	0.792	0.982
			LightGBM	0.989	0.958	0.839	0.958	0.717	0.769	0.978
			CatBoost	0.995	0.972	0.881	0.972	0.786	0.818	0.985
PC + FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.986	0.972	0.881	0.972	0.786	0.818	0.985
			LightGBM	0.994	0.972	0.881	0.972	0.786	0.818	0.985
			CatBoost	0.985	0.972	0.881	0.972	0.786	0.818	0.985
PC + FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.986	0.958	0.839	0.958	0.717	0.769	0.978
			LightGBM	0.979	0.958	0.839	0.958	0.717	0.769	0.978
			CatBoost	0.991	0.958	0.839	0.958	0.717	0.769	0.978
PC + FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.979	0.965	0.859	0.965	0.750	0.792	0.982
			LightGBM	0.982	0.965	0.859	0.965	0.750	0.792	0.982
			CatBoost	0.983	0.965	0.859	0.965	0.750	0.792	0.982

4.3.3. Key Observations on Precision-Recall Trade-Offs

The ClinTox toxicity prediction task exhibits distinct precision-recall dynamics:

- High recall is consistently achieved across most methods (typically > 0.90), reflecting the ability of the models to identify toxic compounds accurately.
- PC + Base MolFormer with LightGBM achieves the best balance with highest MCC (0.875) and near-perfect Recall (0.993) while maintaining strong Precision (0.889).
- End-to-end LoRA methods show very high recall (0.978–0.985) but lower precision (0.769–0.818), suggesting potential over-prediction of toxicity which can be attributed to huge class-imabalance in the dataset.
- Base MolFormer with LightGBM achieves perfectly balanced Precision (0.925) and Recall (0.925), representing an ideal trade-off.

4.3.4. Unique Characteristics of the ClinTox Task

Unlike the BBBP, Flavor and BACE tasks where E2E LoRA finetuning achieves the best results, the ClinTox task shows that:

- Base embeddings are highly effective for toxicity prediction, suggesting that pretrained CLMs already capture relevant toxicity-related molecular features.
- Finetuning provides no additional benefit and may slightly degrade performance, possibly due to the relatively small dataset size or the task's alignment with pretraining objectives.
- Combining PC features with base embeddings yields optimal results, indicating that traditional molecular descriptors provide complementary information for toxicity prediction.
- MolFormer consistently outperforms ChemBERTa across all experimental conditions, reinforcing its superiority as a molecular foundation model (see Tables S6 and S7).

These results demonstrate that the optimal approach for molecular property prediction is task-dependent, and practitioners should evaluate both finetuned and base embedding approaches when developing toxicity prediction models.

4.4. BACE Task Performance

We perform a comprehensive set of comparisons for the BACE dataset.

4.4.1. Top Performing Methods

Our comprehensive evaluation on the BACE dataset reveals that end-to-end LoRA finetuning of MolFormer yields the strongest overall performance. The BACE task presents a more challenging classification problem compared to other datasets, with overall lower performance metrics across all methods as observed from Tables S8 and S9. The top three methods based on average performance across all evaluation metrics are:

- E2E LoRA MolFormer-XL-10pct-both (Focal Loss)-End-to-end LoRA finetuning of MolFormer with focal loss achieved the best overall performance, with an MCC of 0.577 and F1 macro of 0.785. This method attained the highest Accuracy of 0.789, with Precision of 0.783 and Recall of 0.794, demonstrating the most balanced and robust classification performance for BACE inhibitor prediction.
- PC + Base MolFormer-XL-10pct-both with CatBoost - Combining physico-chemical features with base MolFormer embeddings using CatBoost ranked second, achieving an MCC of 0.557 and F1 macro of 0.762. The method attained an AUC of 0.871 and Accuracy of 0.763, with Precision of 0.773 and Recall of 0.784, demonstrating that the combination of engineered features with pretrained representations provides strong performance even without task-specific finetuning.
- FT Embed MolFormer-XL-10pct-both (Focal Loss) with CatBoost - Finetuned MolFormer embeddings combined with CatBoost using focal loss ranked third, achieving an MCC of 0.557 and F1 macro of 0.762. With an AUC of 0.869 and Accuracy of 0.763, this method achieved Precision of 0.773 and Recall of 0.784, confirming the effectiveness of task-specific embedding adaptation.

4.4.2. Performance Trends Across Feature Modalities

A clear trend of performance improvement emerges when progressing from simpler to more sophisticated feature representations as depicted in Tables S8 and S9.

- Physico-chemical (PC) features alone: The baseline using only PC features with tree-based models achieved moderate performance. XGBoost performed best with MCC of 0.490, F1 macro of 0.723, Precision of 0.742, and Recall of 0.749. LightGBM and CatBoost showed slightly lower performance with MCC of 0.463 and 0.443 respectively, indicating the challenging nature of this prediction task.
- Base CLM embeddings: Zero-shot embeddings from pretrained CLMs showed variable results. ChemBERTa-77M MLM with LightGBM achieved MCC of 0.461, F1 macro of 0.704, Precision of 0.729, and Recall of

0.732. ChemBERTa-77M MTR with LightGBM showed improved performance with MCC of 0.514, F1 macro of 0.717, and notably higher Precision (0.760) and Recall (0.755). MolFormer-XL with CatBoost achieved the best base embedding performance with MCC of 0.543, F1 macro of 0.723, Precision of 0.777, and Recall of 0.766, substantially outperforming PC features alone.

- PC + Base embeddings: Combining PC features with base embeddings yielded substantial improvements and produced one of the top three performing methods. PC + ChemBERTa-77M MLM with LightGBM achieved MCC of 0.508, F1 macro of 0.730, Precision of 0.751, and Recall of 0.757. PC + ChemBERTa-77M MTR with LightGBM reached MCC of 0.516, F1 macro of 0.730, with Precision of 0.756 and Recall of 0.760. PC + MolFormer with CatBoost achieved the best combined base performance with MCC of 0.557, F1 macro of 0.762, Precision of 0.773, and Recall of 0.784, ranking as the second-best method overall.
- Finetuned (FT) embeddings: Task-specific finetuning substantially improved performance for MolFormer-based approaches. FT MolFormer embeddings with CatBoost (Focal Loss) achieved MCC of 0.557, F1 macro of 0.762, Precision of 0.773, and Recall of 0.784. FT MolFormer with LightGBM (Weighted Loss) reached MCC of 0.550, F1 macro of 0.762, Precision of 0.769, and Recall of 0.781. However, FT ChemBERTa embeddings showed limited improvement, with best performance (Weighted Loss + CatBoost) achieving only MCC of 0.480 and F1 macro of 0.717.
- PC + FT embeddings: The combination of PC features with finetuned embeddings did not outperform finetuned embeddings alone. PC + FT MolFormer (Weighted Loss) with LightGBM achieved MCC of 0.550, F1 macro of 0.762, Precision of 0.769, and Recall of 0.781, matching but not exceeding FT embeddings alone. PC + FT ChemBERTa combinations generally underperformed, with MCC ranging from 0.382 to 0.480.
- End-to-end LoRA finetuning: Direct optimization through LoRA achieved the best performance overall (see Table S8). E2E LoRA MolFormer (Focal Loss) attained the best MCC of 0.577, F1 macro of 0.785, Accuracy of 0.789, with well-balanced Precision (0.783) and Recall (0.794). E2E LoRA MolFormer (Weighted Loss) showed lower performance with MCC of 0.476 and F1 macro of 0.722. Notably, E2E LoRA ChemBERTa-77M MLM (Weighted Loss) achieved competitive MCC of 0.510 and F1 macro of 0.736, outperforming its focal loss variant (MCC: 0.384, F1 macro: 0.651) but still significantly lower than base embeddings.

4.4.3. Key Observations on Precision-Recall Trade-Offs

The BACE task exhibits relatively balanced precision-recall dynamics across methods:

- E2E LoRA MolFormer (Focal Loss) achieves the best balance with Precision of 0.783 and Recall of 0.794, indicating robust identification of BACE inhibitors.
- Base MolFormer with CatBoost showed high Precision (0.777) relative to Recall (0.766), suggesting conservative predictions.
- PC features alone demonstrate balanced but lower Precision (0.742) and Recall (0.749) with XGBoost.
- Finetuned embeddings generally improve both precision and recall compared to base embeddings, with FT MolFormer + CatBoost with focal loss achieving Precision of 0.773 and Recall of 0.784.
- Focal loss consistently outperforms weighted loss for MolFormer-based approaches on this task, particularly for end-to-end LoRA finetuning.

4.4.4. Unique Characteristics of the BACE Task

The BACE task presents several distinct characteristics:

- Most challenging prediction task among all the evaluated datasets, with the highest-performing method achieving only MCC of 0.577 compared to 0.797 (BBBP), 0.801 (Flavor), and 0.875 (ClinTox).
- End-to-end LoRA finetuning with focal loss is optimal, aligning with BBBP and Flavor tasks but contrasting with ClinTox where base embeddings performed best.
- MolFormer substantially outperforms ChemBERTa across all conditions, with performance gaps more pronounced than in other tasks.
- Focal loss outperforms weighted loss for MolFormer-based methods, suggesting that focusing on hard-to-classify examples is particularly beneficial for this challenging task.
- PC + Base embeddings achieve top-3 performance, indicating that combining engineered features with zero-shot pretrained representations remains a competitive approach even without finetuning.
- Adding PC features to finetuned embeddings provides minimal additional benefit, indicating that the finetuned neural representations already capture the relevant molecular information for BACE inhibition prediction.

These results demonstrate that for challenging molecular property prediction tasks like BACE inhibitor

classification, E2E LoRA finetuning with focal loss provides the optimal approach, and the choice of base CLM (MolFormer vs. ChemBERTa) significantly impacts performance.

Table S8. Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results are highlighted in bold.

Modality	CLM	ML-Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
PC	-	XGBoost	0.837	0.724	0.723	0.724	0.490	0.742	0.749
PC	-	LightGBM	0.830	0.711	0.710	0.711	0.463	0.728	0.735
PC	-	CatBoost	0.838	0.697	0.697	0.697	0.443	0.720	0.724
Base	ChemBERTa-77M MLM	XGBoost	0.811	0.684	0.684	0.684	0.424	0.711	0.713
		LightGBM	0.817	0.704	0.704	0.704	0.461	0.729	0.732
		CatBoost	0.800	0.704	0.704	0.704	0.453	0.724	0.729
Base	ChemBERTa-77M MTR	XGBoost	0.830	0.697	0.697	0.697	0.468	0.736	0.733
		LightGBM	0.828	0.717	0.717	0.717	0.514	0.760	0.755
		CatBoost	0.831	0.671	0.670	0.671	0.451	0.735	0.717
Base	MolFormer-XL-10pct-both	XGBoost	0.872	0.684	0.684	0.684	0.449	0.728	0.722
		LightGBM	0.863	0.691	0.690	0.691	0.478	0.745	0.733
		CatBoost	**0.885**	0.724	0.723	0.724	0.543	**0.777**	0.766
PC + Base	ChemBERTa-77M MLM	XGBoost	0.827	0.717	0.717	0.717	0.473	0.733	0.740
		LightGBM	0.841	0.730	0.730	0.730	0.508	0.751	0.757
		CatBoost	0.842	0.697	0.697	0.697	0.443	0.720	0.724
PC + Base	ChemBERTa-77M MTR	XGBoost	0.845	0.724	0.723	0.724	0.490	0.742	0.749
		LightGBM	0.845	0.730	0.730	0.730	0.516	0.756	0.760
		CatBoost	0.851	0.704	0.704	0.704	0.487	0.747	0.741
PC + Base	MolFormer-XL-10pct-both	XGBoost	0.862	0.724	0.723	0.724	0.490	0.742	0.749
		LightGBM	0.865	0.697	0.697	0.697	0.451	0.724	0.727
		CatBoost	0.871	**0.763**	**0.762**	**0.763**	**0.557**	0.773	**0.784**

Table S9. Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold.

Modality	CLM	Loss Function	ML Model	AUC	Accuracy	F1 macro	F1 micro	MCC	Precision	Recall
E2E LoRA	ChemBERTa-77M MLM	Focal Loss	-	0.812	0.651	0.651	0.651	0.384	0.695	0.689
		Weighted Loss	-	0.850	0.737	0.736	0.737	0.510	0.750	0.759
E2E LoRA	ChemBERTa-77M MTR	Focal Loss	-	0.846	0.724	0.721	0.724	0.463	0.727	0.737
		Weighted Loss	-	0.815	0.724	0.721	0.724	0.463	0.727	0.737
E2E LoRA	MolFormer-XL-10pct-both	Focal Loss	-	**0.872**	**0.789**	**0.785**	**0.789**	**0.577**	**0.783**	**0.794**
		Weighted Loss	-	0.871	0.724	0.722	0.724	0.476	0.733	0.743
FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.825	0.671	0.671	0.671	0.404	0.702	0.702
			LightGBM	0.811	0.671	0.671	0.671	0.389	0.692	0.696
			CatBoost	0.823	0.678	0.678	0.678	0.414	0.706	0.708
FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.841	0.711	0.710	0.711	0.471	0.733	0.738
			LightGBM	0.835	0.711	0.710	0.711	0.471	0.733	0.738
			CatBoost	0.851	0.717	0.717	0.717	0.480	0.737	0.743
FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.845	0.684	0.684	0.684	0.416	0.706	0.710
			LightGBM	0.845	0.717	0.716	0.717	0.459	0.725	0.734
			CatBoost	0.847	0.711	0.709	0.711	0.449	0.720	0.729
FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.792	0.697	0.697	0.697	0.429	0.711	0.718
			LightGBM	0.820	0.671	0.671	0.671	0.382	0.688	0.693
			CatBoost	0.818	0.691	0.690	0.691	0.419	0.706	0.713
FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.866	0.750	0.749	0.750	0.537	0.764	0.773
			LightGBM	0.869	0.737	0.736	0.737	0.517	0.755	0.762
			CatBoost	0.869	0.763	0.762	0.763	0.557	0.773	0.784
FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.856	0.757	0.755	0.757	0.540	0.765	0.776
			LightGBM	0.860	0.763	0.762	0.763	0.550	0.769	0.781
			CatBoost	0.848	0.757	0.755	0.757	0.540	0.765	0.776
PC + FT Embed	ChemBERTa-77 MLM	Focal Loss	XGBoost	0.819	0.671	0.671	0.671	0.396	0.697	0.699
			LightGBM	0.819	0.664	0.664	0.664	0.395	0.698	0.697
			CatBoost	0.822	0.684	0.684	0.684	0.424	0.711	0.713
PC + FT Embed	ChemBERTa-77 MLM	Weighted Loss	XGBoost	0.854	0.704	0.704	0.704	0.461	0.729	0.732
			LightGBM	0.829	0.717	0.717	0.717	0.480	0.737	0.743
			CatBoost	0.849	0.717	0.717	0.717	0.480	0.737	0.743
PC + FT Embed	ChemBERTa-77 MTR	Focal Loss	XGBoost	0.844	0.704	0.703	0.704	0.439	0.715	0.724
			LightGBM	0.849	0.697	0.697	0.697	0.429	0.711	0.718
			CatBoost	0.853	0.711	0.709	0.711	0.449	0.720	0.729
PC + FT Embed	ChemBERTa-77 MTR	Weighted Loss	XGBoost	0.792	0.691	0.691	0.691	0.434	0.715	0.718
			LightGBM	0.804	0.684	0.684	0.684	0.409	0.702	0.707
			CatBoost	0.819	0.691	0.690	0.691	0.419	0.706	0.713
PC + FT Embed	MolFormer-XL-10pct-both	Focal Loss	XGBoost	0.870	0.737	0.736	0.737	0.517	0.755	0.762
			LightGBM	0.869	0.717	0.717	0.717	0.473	0.733	0.740
			CatBoost	0.866	0.757	0.755	0.757	0.540	0.765	0.776
PC + FT Embed	MolFormer-XL-10pct-both	Weighted Loss	XGBoost	0.856	0.750	0.748	0.750	0.524	0.756	0.767
			LightGBM	0.857	0.763	0.762	0.763	0.550	0.769	0.781
			CatBoost	0.839	0.757	0.755	0.757	0.534	0.761	0.773

References

1. Nowakowska, S. ChemBERTa-2: Fine-Tuning for Molecule's HIV Replication Inhibition Prediction. *ChemRxiv* **2023**. https://doi.org/10.26434/chemrxiv-2023-b57vx.
2. Lee, D.; Cho, Y. Fine-Tuning Pocket-Conditioned 3D Molecule Generation via Reinforcement Learning. In Proceedings of the ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design, Vienna, Austria, 7–11 May 2024.
3. Maleki, S.; Huetter, J.C.; Chuang, K.V.; et al. Efficient Fine-Tuning of Single-Cell Foundation Models Enables Zero-Shot Molecular Perturbation Prediction. *arXiv* **2024**, arXiv:2412.13478.
4. Ren, W.; Li, X.; Wang, L.; et al. Analyzing and Reducing Catastrophic Forgetting in Parameter Efficient Tuning. *arXiv* **2024**, arXiv:2402.18865.
5. Schuh, M.G.; Boldini, D.; Sieber, S.A. Synergizing Chemical Structures and Bioassay Descriptions for Enhanced Molecular Property Prediction in Drug Discovery. *J. Chem. Inf. Model.* **2024**, *64*, 4640–4650.
6. Hu, E.J.; Shen, Y.; Wallis, P.; et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR* **2022**, *1*, 3.
7. Balne, C.C.S.; Bhaduri, S.; Roy, T.; et al. Parameter Efficient Fine Tuning: A Comprehensive Analysis Across Applications. *arXiv* **2024**, arXiv:2404.13506.